General and specific factors in the processing of faces.

Roeland J Verhallen[a], Jenny M Bosten[a,b], Patrick T Goodbourn[a,c], Adam J Lawrance-Owen[a], Gary Bargary[a,d] & J D Mollon[a]

[a] Department of Psychology, Downing St., Cambridge CB2 3EB, United Kingdom

[b] School of Psychology, University of Sussex, Brighton, UK

[c] School of Psychological Sciences, University of Melbourne, Australia

[d] Division of Optometry and Visual Science, City University London, United Kingdom

CORRESPONDING AUTHOR

Roeland J Verhallen (rjv31@cam.ac.uk)

Department of Psychology,

Downing St.,

Cambridge CB2 3EB,

United Kingdom

+44 1223 333580

ABSTRACT

The ability to recognize faces varies considerably between individuals, but does performance co-vary for tests of different aspects of face processing? For 397 participants (of whom the majority were university students) we obtained scores on the Mooney Face Test, Glasgow Face Matching Test (GFMT), Cambridge Face Memory Test (CFMT) and Composite Face Test. Overall performance was significantly correlated for each pair of tests, and we suggest the term *f* for the factor underlying this pattern of positive correlations. However, there were large variations in the amount of variance shared by individual tests: The GFMT and CFMT are strongly related, whereas the GFMT and the Mooney test tap largely independent abilities. We do not replicate a frequently reported relationship between holistic processing (from the Composite test) and face recognition (from the CFMT)—indeed, holistic processing does not correlate with any of our tests. We report associations of performance with digit ratio and autism-spectrum quotient (AQ), and from our genome-wide association study we include a list of suggestive genetic associations with performance on the four face tests, as well as with *f*.

KEYWORDS

Face perception; face recognition; f; holistic processing; individual differences; autism-spectrum quotient (AQ); genome-wide association study (GWAS)

1. INTRODUCTION

Face recognition is singularly important for human social interaction (Bruce and Young, 2012), but not everyone is equally good at recognizing faces. Indeed, there are large individual differences: Some people cannot recognize faces at all, while others remember practically every face they see (Burton et al., 2010; Duchaine et al, 2007; Russell et al., 2009). In some situations, quantifying the ability to detect, discriminate and recognize faces is of great practical value—for example, in the screening of border-control officers (Burton and Jenkins, 2011). However, in the history of understanding perceptual and cognitive processes, the measurement of individual differences has led also to theoretical insights. Thus Peterzell and Teller (2000) used a covariance analysis to identify sub-channels within the visual system that are specific to particular spatial frequency bands; and in the specific case of face processing, studies of individual differences have shown that there is remarkably little overlap between general intelligence and the specific ability to recognize faces (Wilmer et al., 2014; Shakeshaft & Plomin, 2015)

Several tests have been developed to measure the ability to detect faces or to remember them, but no single test assesses all aspects of face processing. We here ask to what extent different measures co-vary. For a large sample of healthy participants, we established the distribution of individual performance on four well-established tests of face processing: The Mooney Face Test, the Glasgow Face Matching Test, the Cambridge Face Memory Test, and the Composite Face Test.

The stimuli of the classical *Mooney Face Test* consist of seemingly unrelated patches of pure black and pure white, which, without apparent conscious effort on the viewer's part, suddenly arrange themselves to form the percept of a face (Mooney, 1957a, 1957b). This process of organization is referred to as *closure*. The objective of the Mooney test is to detect the face, and the test is considered a test of face detection and of *holistic processing*—the processing of faces as a whole as opposed to processing of individual features separately.

The *Glasgow Face Matching Test* (GFMT) measures discrimination between unfamiliar faces. Participants are shown two photographs of faces and asked to indicate whether they are of the same person, or of different persons (Burton et al., 2010). Contrary to intuition, performance is far from perfect and there are marked individual differences.

The *Cambridge Face Memory Test* (CFMT) is widely used to assess face recognition

ability and is often administered via the Internet (Duchaine and Nakayama, 2006; Wilmer et al., 2010). Individuals with prosopagnosia show significantly lower performance than controls (Duchaine and Nakayama, 2006), and performance is highly heritable (Wilmer et al., 2010).

The *Composite Face Test* is often-used but unstandardized: Many researchers have created their own version (Richler et al., 2011; Richler and Gauthier, 2014; Rossion, 2013; Young, Hellawell and Hay, 1987). In the Composite test, the participant makes a same/different judgment between the top half of the 'study' face and the top half of the subsequently presented 'target' face, while ignoring the bottom halves. Face stimuli are created by combining a top half and a bottom half, either of the same face or of different faces; the two halves are either aligned or misaligned. On a given trial, both—or either—the top and the bottom half of each face may differ between the study face and the target face, or may be the same. The test is designed to tap into holistic processing: The bottom half should influence the perception of the top half in the aligned conditions, since then all the features cohere in a Gestalt; and if the top halves are the same but the bottom halves differ, this holistic process would interfere with making a correct judgment.

All four tests previously have been compared to other tests, though not necessarily to one another. Foreman (1991) tested 127 participants on a visual-search task, the Mooney test, and two other tests of closure (the Gollin Incomplete Figures Test and the Poppelreuter test), but found no significant correlation in performance between the Mooney test and any of the other tests. This suggests that Mooney performance is independent of visual-search efficiency, and that the Mooney test does not tap the same processes as the two other tests of closure.

Burton and colleagues (2010) compared the Glasgow Face Matching Test to three measures of visual processing in a sample of 300 participants. GFMT performance correlated significantly and moderately strongly with matching of familiar line drawings of figures ($r = .42$, $p < .001$), and significantly but less strongly with recognition memory for faces ($r = .29$, $p < .001$). There was no significant correlation with visual short-term memory for objects ($r = .05$, $p > .05$).

Bowles and colleagues (2009) report a significant and strong correlation ($r = –.61$, $p < .001$, $N = 124$) between the Cambridge Face Memory Test and the Cambridge Face Perception Test, which asks participants to sort a row of faces from "most similar" to "most dissimilar" in comparison to a target face; the correlation is negative because the measure of the latter test is the number of errors, rather than number correct, as is the

case for the former). Wilmer and colleagues (2012; 2014) report a significant and sizeable correlation between the CFMT and a Famous Faces Test (r = .52, N = 1,219), but only relatively low correlations between the CFMT and two other memory tests: The Abstract Art Memory Test (r = .26, N = 1,469) and a Verbal Paired-Associates Memory Test (r = .18, N = 1,469). It is on the basis of these—and other—results, that Wilmer and colleagues argue that face recognition is an independent skill, exhibiting high correlations with other tasks of face processing, but low correlations with other abilities, such as general memory.

Several studies have investigated the relationship between face recognition and holistic processing, but results are mixed: Some report a positive correlation—either strong (DeGutis et al., 2013; Richler et al., 2011) or moderate (Wang et al., 2012)—whereas others observe no significant correlation (Konar et al., 2010). The interpretation of these studies is complicated by differences in both methodology and data analysis (DeGutis et al., 2013; Richler and Gauthier, 2014; Rossion, 2013).

In the present study, a large cohort of participants completed four tests that measure different aspects of face processing. The tests were selected to reliably assess as many different aspects of face processing as possible, while keeping our online test battery sufficiently brief as to encourage a high rate of participation and completion. Additionally, we hold genetic and phenotypic data for our participants from their previous visits to our lab. Face recognition previously has been shown to be strongly heritable (Shakeshaft & Plomin, 2015; Wilmer et al., 2010), to be impaired in people with autism (e.g. Weigelt et al., 2012), and to be related to digit ratio (Leow and Davis, 2012). We are in a position to report results from a genome-wide association study (GWAS) that we conducted on participants' performance on our four face tests. We also report results from correlations of performance on our four tests with both autism-spectrum quotient and digit ratio.

2. METHODS

2.1. Participants

Our 397 participants (252 female) were a subset of a cohort of 1,060 who had previously completed a battery of perceptual tests in our laboratory as part of the PERGENIC project (Goodbourn et al., 2012; Lawrance-Owen et al., 2013; Verhallen et al., 2014). Participants were healthy young adults between the ages of 18 and 42 (M = 24 years, SD = 4.3), all of European descent. When tested on their original visit to the laboratory, 97% of the present cohort had a (corrected) visual acuity of 0.2 logMAR or better. The majority were students at the University of Cambridge. Participants took part in order to have a chance of winning a Kindle 3G or Amazon vouchers worth £120, the winner being chosen randomly from all who completed the four tests. Ethical permission for the study was given by the Cambridge University Psychology Ethics Committee, and work was carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Participants gave informed consent before testing began.

2.2. Materials

The Mooney test was classically designed to be administered by personal interview; in the current study we use our online, three-alternative forced-choice (3AFC) version of the Mooney test (Verhallen et al., 2014). The test uses the original forty Mooney (1957a) faces, but each face is paired with two custom-made distractors. The position of the target image was random and 3AFC stimuli remained on screen until participants made a response by pressing the keys 1, 2 or 3 on their keyboard. The first trial, of forty in total, was a practice trial with feedback.

The shortened version of the Glasgow Face Matching Test was administered according to the original procedure (Burton et al., 2010): For forty trials participants had to indicate whether two photographs were of the same person or of different persons, by pressing the keys *L* or *A* on their keyboard, respectively. Each greyscale photograph was cropped tightly around the external outline of the face, ears and hair, and was presented on a white background. Stimuli remained on screen until participants made a response. In line with the original procedure there was no practice trial.

The Cambridge Face Memory Test was administered according to the original procedure (Duchaine and Nakayama, 2006): The first of three sections introduced six

different faces for memorization, each presented for three seconds, followed by three 3AFC tests for each face. Each greyscale photograph was cropped with an oval frame that masks external features (hair and ears), and was presented on a black background. Sections 2 and 3 used these same six faces to test face memory: Participants were shown, for ten seconds at the beginning of each section, all six faces in an array. Sections 2 and 3 were of increased difficulty because of differing lighting and viewing angles between pairs of photographs (section 2) or because of the superposition of noise (section 3). One practice trial with feedback preceded the test.

The Composite Face Test used in this study was the version developed by Richler and colleagues (Richler et al., 2011) incorporating stimuli from the Max Planck Institute Face Database (Troje and Bülthoff, 1996). There is debate about the differential merits of two existing designs, the *partial* design and the *complete* design (Richler and Gauthier, 2013; Rossion, 2013). By using the Composite test from Richler and colleagues we opted for the complete design; we did so because this design also allows us to approximate—after data collection—the measure that the partial design would have yielded. The test consisted of 160 trials in which a greyscale composite face was shown for 200 ms (the study face), followed by a blank inter-stimulus interval of 500 ms and then a target composite face shown for 200 ms. Each face was presented on a black background, cropped tightly around the external outline of the face including the ears, but with the hair and hairline masked. Participants were asked to use their keyboard to indicate whether the top halves of the two faces were the same (*L*-key) or different (*A*-key), while ignoring the lower half. The first trial was a practice trial with feedback.

Each of the 160 trials in the Composite test is categorized on three variables: 1. "Similarity:" Whether the top halves of the study and target faces are the *same* or *different* (this judgment constitutes the task of the participant); 2. "Alignment:" Whether the top and bottom halves of the target face are *aligned* or *misaligned* (the study face was always aligned); and 3. "Congruency:" Whether the similarity of the bottom halves between the study and target faces is *congruent* or *incongruent* with the similarity of their top halves.

The measure of interest for the Composite test is not the overall score for all these conditions, but rather the *holistic index* (Richler et al., 2011). First, a specific combination of the four conditions (see Supplementary Materials, S.1, for detailed calculations) is used to calculate two variables: The *condition of interest* (i.e. *aligned congruent* trials minus *aligned incongruent* trials), and the *control condition* (i.e. *misaligned congruent* trials minus *misaligned incongruent* trials). Then, the residuals taken from regressing the variation

of the *control condition* out of the *condition of interest* constitute the *holistic index* (DeGutis et al., 2013). The rationale is that—for the aligned trials—participants who have strong holistic processing will experience high interference from the bottom half of the face: If the change in bottom halves is congruent with the change in top halves, then these participants' performance is aided, but if the change in bottom halves is incongruent with the change in top halves, performance is impaired. For the misaligned trials, the assumption is that misalignment breaks holistic processing, since the faces no longer form a coherent whole; the misaligned condition is thus not a measure of holistic processing.

2.3. Procedure

The present data were collected online, although all the participants were personally known to us from their previous visits to the laboratory. All 1,060 participants of the original cohort were sent a web-link to the online tests; 397 of them completed all four tests. Each of these 397 participants completed the four tests in the same sequence: The modified Mooney Face Test, the Glasgow Face Matching Test, the Cambridge Face Memory Test, and the Composite Face Test. No feedback was given for any test, except for practice trials as indicated previously. Participants were instructed to respond as quickly and as accurately as possible; their response times were recorded, though not restricted. Before beginning the tests, participants subjectively rated their face recognition ability in response to the question "On a scale of 1 to 10 (with 1 being really bad, and 10 being really good), where would you place yourself in terms of recognizing faces?" Data analysis was performed using R, unless indicated otherwise.

3. RESULTS

3.1. Distributions and correlations for the four tests of face processing

The range of scores is wide for all tests. The mean proportion correct for the modified Mooney Face Test is 34.9 trials out of 39 (*SD* = 2.8, range 25 to 39; 30 participants at ceiling), for the Glasgow Face Matching Test 31.5 trials out of 40 (*SD* = 4.6, range 14 to 40; four participants at ceiling, five participants at or below chance level), for the Cambridge Face Memory Test 54.3 trials out of 72 (SD = 9.1, range 26 to 72; one participant at ceiling), and for the Composite Face Test 137.8 trials out of 160 (*SD* = 11.6, range 79 to 157; one participant below chance level); as the *holistic index* is a standardized residual, it has a mean of 0 and *SD* of 1.0 (range −2.62 to 3.53; see Table 2 for further statistics). To allow comparison of the raw scores of the different tests, we give in Table 1 the performance scores converted to percentages.

TABLE 1

Summary Statistics for the four Tests. The Minimum (*Min.*), Mean and Maximum (*Max.*) Scores in Percentages, Standard Deviation (*SD*) in Percentage, Chance of guessing correctly (*Chance*), and Guttman's Reliability Indices λ_2, λ_3 (i.e. Cronbach's alpha) and λ_6 for our four Tests: the Modified Mooney Face Test (*Mooney*), the Glasgow Face Matching Test (*GFMT*), the Cambridge Face Memory Test (*CFMT*), and Overall Raw Score of the Composite Face Test (*Comp – Raw Score*), including Minimum, Mean, Maximum, and Standard Deviation, as well as λ_3, for the Holistic Index (*Holistic*) of the Composite Face Test.

	Min.	**Mean**	**Max.**	**SD**	**Chance**	λ_2	λ_3	λ_6
Mooney	64	90	100	7.2	⅓	.69	.67	.69
GFMT	35	79	100	11.5	½	.72	.71	.76
CFMT	36	75	100	12.6	⅓	.89	.88	.91
Composite (*Raw Score*)	49	86	98	7.3	½	.88	.88	.94
Holistic Index	−2.62	0	3.53	1.0	*n.a.*	*n.a* [1]	.53	*n.a.*

For the Mooney test, our sample's results are comparable to those reported by Vigen and colleagues (1982), who—for a sample of 100 undergraduates—find a mean performance of 81.0% correct (*SD* = 6.6%) using the Mooney stimuli in a lab-based experiment. Our participants' mean score and range of performance for the GFMT are comparable to previously reported results (Burton et al., 2010: *M* = 81%, *SD* = 9.7%), though our distribution extends slightly further at the lower end. Performance on the CFMT is also comparable to previous studies (Bowles et al., 2009; Wilmer et al., 2010),

[1] Guttman's λ_2 and λ_6 both require raw data, and thus cannot be calculated for the holistic index, which uses *d'*. Instead, we manually calculated split-half reliability, the result of which we report in the λ_3 column: a Spearman-Brown corrected reliability of ρ = .53 (*SD* = .06), the mean of 5,000 splits of the data. See Supplementary Materials (S.2) for details.

and overlaps at the lower end with the range of performance by individuals with prosopagnosia (Duchaine and Nakayama, 2006). Moreover, the correlations we observe when comparing performance of the three parts of the CFMT to one another are almost identical to those observed by Duchaine & Nakayama (2006): We observe Spearman's correlations of $\rho = .34$ between parts 1 and 2, $\rho = .41$ between parts 1 and 3, and $\rho = .74$ between parts 2 and 3.

When we investigate plots from DeGutis et al., (2013, their Figure 4C), our distribution of the holistic index seems similar, though wider; it exhibits kurtosis of .60 and a slight positive skew of .29 (see also Table 2 for the distributions of d' broken down by condition; and see Figure 2 in §3.5 for a plot of d' broken down by the conditions *alignment* and *congruency*).

TABLE 2

Summary Statistics for the four Conditions of the Composite Face Test. The Minimum (*Min.*), Mean, Maximum (*Max.*), and Standard Deviation (*SD*) of d', as well as the Percentage of Participants who were at Ceiling, and the Kurtosis and Skew of the Distribution of d', separately for the four Conditions *Aligned Congruent*, *Aligned Incongruent*, *Misaligned Congruent*, and *Misaligned Incongruent*.

	Min.	**Mean**	**Max.**	**SD**	**% Ceiling**	**Kurtosis**	**Skew**
Aligned Congruent	.12	3.02	3.96	.68	16.4	.74	-.69
Aligned Incongruent	-3.16	1.63	3.96	.94	.8	2.61	-.58
Misaligned Congruent	.00	2.39	3.96	.71	3.0	.10	-.11
Misaligned Incongruent	-1.82	2.30	3.96	.86	3.5	2.31	-.71

Correlations between performances on each pair of tests are highly significant (see Table 3), except for pairs that included the Composite Face Test's *holistic index* (for which *p*-values ranged between .04 and .53, before Bonferroni correction). However, when we simply consider the raw score on the Composite test (the number of trials to which a participant responded correctly) we do observe significant correlations with performance on each of the other three tests.

Since the distributions of scores were not normal, we give both Pearson's *r* and Spearman's ρ; the corresponding values are similar. For the four tests, the shared variance of the significant inter-correlations (estimated from the square of Pearson's *r*) ranges from a fairly high 23% between the Cambridge Face Memory Test and the Glasgow Face Matching Test (see Figure 1A), to a low 4% between the Mooney Face Test and the Glasgow Face Matching Test (see Figure 1B).

TABLE 3

Correlations between Performance on Pairs of Tests: Pearson's r and Spearman's ρ for all Combinations of the four Tests: The modified Mooney Face Test (*Mooney*), the Glasgow Face Matching Test (*GFMT*), the Cambridge Face Memory Test (*CFMT*), and the Composite Face Test's Holistic Index (*Holistic; d'*), as well as the Composite Face Test's overall Raw Score (*Raw Score*). All correlations use the full sample size of $N = 397$, and p-values are uncorrected. Confidence intervals at 95% are given between square brackets.

	Mooney		**GFMT**		**CFMT**		**Holistic**	
	Pearson	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*
GFMT	.20 ** [.10, .29]	.21 ** [.11, .30]						
CFMT	.31 ** [.22, .39]	.31 ** [.22, .40]	.48 ** [.40, .55]	.49 ** [.41, .56]				
Holistic	–.06 n.s. [–.16, .03]	–.09 n.s. [–.19, .01]	–.02 n.s. [–.11, .08]	–.01 n.s. [–.11, .09]	–.02 n.s. [–.12, .08]	–.03 n.s. [–.13, .07]		
Raw Score	.19 * [.09, .28]	.20 ** [.10, .29]	.26 ** [.17, .35]	.33 ** [.24, .41]	.40 ** [.31, .48]	.42 ** [.34, .50]	–.30 ** [–.39, –.21]	–.26 ** [–.35, –.17]

* $p < .001$

** $p \ll .0001$

n.s. = not significant

To judge whether all trials of each of the four tests that we used were informative, we investigate performance per item, for each test. This item analysis shows that, for the GFMT and the Composite Face Test, no item is solved by all participants, whereas for the modified Mooney Face Test two items are solved by all participants (items 7 & 17), and for the CFMT, one (item 1). Participants perform below chance level on one item in the GFMT (item 27) and on one item in the Composite Face Test (item 3). In Table 1 we also report the internal reliabilities of the four tests calculated using Guttman's λ_6 (Guttman, 1945; Revelle and Zinbarg, 2009); we also report Guttman's λ_2 and λ_3 (i.e. Cronbach's alpha) to enable comparison with other studies that report them. Since the calculation of Guttman's λ_2 and λ_6 requires raw performance data, we manually calculate the Spearman-Brown corrected split-half reliability of our holistic index, which we also report in Table 1 (see Supplementary Materials, S.2, for details).

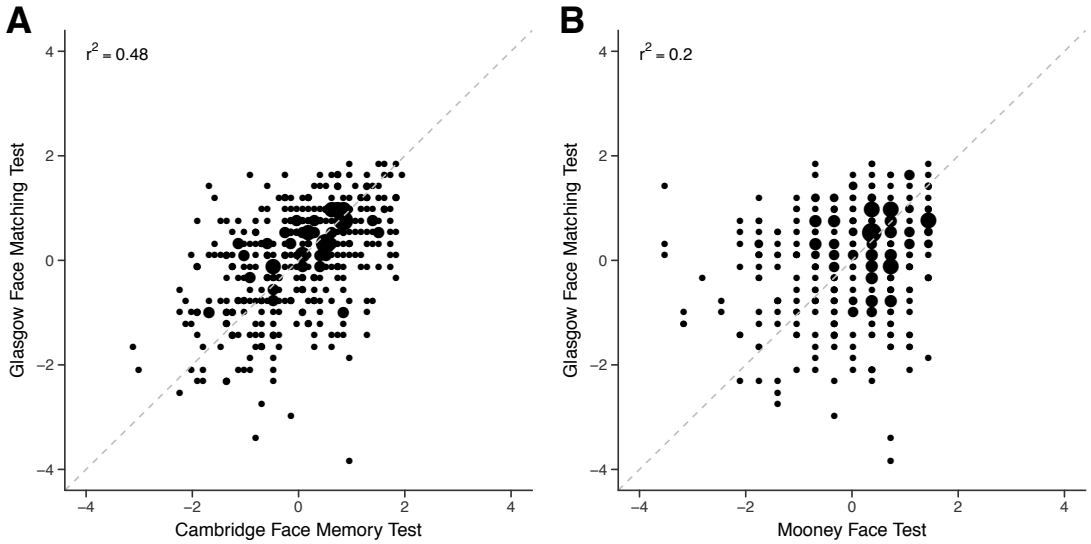

Figure 1. Scatterplots comparing normalised performance (z-scores) across two tests. Panel A shows the two tests with the highest correlation: The Cambridge Face Memory Test and the Glasgow Face Matching Test (Pearson's $r = .48$). Panel B shows the two tests with the lowest correlation: The modified Mooney Face Test and the Glasgow Face Matching Test (Pearson's $r = .20$). Point size is scaled linearly to reflect the number of participants with that particular combination of scores. To aid interpretation on these normalized axes, we include the dashed, grey line ($x = y$), upon which all points would fall in the case of a perfect correlation.

To investigate whether a task is generally performed instantaneously, or rather benefits from longer exposure times, we correlate performance on our four tests with the amount of time taken for each test. We observe a significant correlation between time taken and performance for the Glasgow Face Matching Test only: Participants who took longer tended to have a higher score, although only 6% of variance in accuracy could be predicted from speed (Spearman's $\rho = .23$ [.14, .33], $r^2 = .06$, $p = 2.3 \times 10^{-6}$).

3.2. A common factor underlying performance on tests of face processing: *f*

Although the several tests vary in the extent that they correlate with one another (Table 3), all pairs of measures (except those including the holistic index) do exhibit positive correlations, much in the way that the very diverse subtests of the Wechsler Adult Intelligence Scale exhibit a pattern of positive correlations. We therefore conducted a factor analysis on scores from the four tests, excluding the holistic index. As our non-normally distributed scores may violate assumptions of normally distributed residuals, we applied a *rank-based inverse normal transformation* by which scores are converted to rank orders, with each quantile of the resulting distribution mapped on to the corresponding quantile of a normal distribution. We also included four non-face measures of visual perception from the PERGENIC test battery: 'contrast sensitivity',

i.e. thresholds for detecting sinusoidal gratings of 3 cycles per degree; 'coherent form (sine wave)', i.e. thresholds for detecting the orientation of sinusoidal gratings formed by sinusoidally varying dot density; 'coherent form (Glass patterns)' i.e. thresholds for detecting the orientation of gratings formed by Glass patterns of varying coherence; and 'coherent motion', i.e. percentage of coherent dots needed to report direction in an array of moving dots (for methods see Goodbourn et al., 2012 and Bosten et al., 2015). We selected these measures from the larger battery because they test a range of detection and integration processes that might or might not share variance with different components of face processing. We used SPSS version 21 for the factor analysis. The method of extraction was PCA, and we applied a Varimax rotation. We held data on all measures entered into the factor analysis for 376 of our sample of 397 participants.

The factor analysis identified three factors (by inspection of the Scree plot) that explained a cumulative variance of 61.6% (29.5, 20.4 and 11.7% respectively), and which had Eigenvalues of 2.4, 1.6 and .9, respectively. The first factor loaded positively and strongly on the four face measures (see Table 4), but not on the other measures of form and motion perception. The second factor loaded strongly and positively on the Mooney test, on 'coherent form (Glass patterns)', and on coherent form (sine wave). The third factor loaded strongly and positively on contrast sensitivity and on coherent motion. Table 4 gives the loadings of the three factors with Varimax rotation; but the unrotated factors gave similar results.

The first factor of Table 4 recalls the celebrated factor *g* or 'general ability,' which Spearman judged to underlie all measures of intelligence (Spearman, 1927). We assess its status in the Discussion below, but for the remainder of the Results refer to it as '*f*.'

TABLE 4

For each of the Four Face Tests, as well as for Four Non-Face Measures of Visual Perception, are listed the Loadings of the Three Factors extracted using Factor Analysis. The Loadings given here are the Result of a Varimax Rotation with Kaiser Normalization. For our factor analysis, *N* = 376. For clarity, factor loadings greater than .40 are highlighted in boldface.

Test	Factor 1	Factor 2	Factor 3
Mooney	**.42**	**.51**	–.29
GFMT	**.76**	.02	–.02
CFMT	**.80**	.19	.00
Composite *(Raw Score)*	**.72**	–.03	.20
Coherent form (Glass patterns)	.09	**.82**	.20
Coherent form (sine wave)	–.06	**.76**	.35

TABLE 4

For each of the Four Face Tests, as well as for Four Non-Face Measures of Visual Perception, are listed the Loadings of the Three Factors extracted using Factor Analysis. The Loadings given here are the Result of a Varimax Rotation with Kaiser Normalization. For our factor analysis, N = 376. For clarity, factor loadings greater than .40 are highlighted in boldface.

Test	Factor 1	Factor 2	Factor 3
Contrast sensitivity	.14	.21	**.68**
Coherent motion	–.01	.11	**.80**

3.3 Phenotypic correlates of face-processing ability

Since our participants had previously visited our lab as part of the PERGENIC project, we hold detailed genotypic and phenotypic data for most of them.

Sex & Age. We have previously reported the significant sex difference that we observe for performance on the modified Mooney test (Verhallen et al., 2014); we do not observe a significant sex difference in performance for any of the other tests, nor for *f* (Cohen's *d* ranged from .02 to .31). We observe a significant effect of age for performance on the GFMT (Spearman's ρ = .15 [.05, .24], p = .003), for performance on the CFMT (ρ = .20 [.11, .29], $p = 5.2 \times 10^{-5}$) and for *f* (ρ = .21 [.11, .30], $p = 5.9 \times 10^{-5}$); in the case of the other measures, Spearman's ρ ranged from –.04 to .07.

Self-ratings of facial recognition. Subjective rating of the ability to recognize faces is significantly correlated with performance on all four of the face tests (including overall raw score on the Composite Face Test) and with *f*, but not with the holistic index (see Table 5). On average, participants rate themselves 6.5 on a scale of 1 to 10 (SD = 1.8) with a range covering the full scale.

TABLE 5

The Spearman Correlation Coefficients and Probability Values (after Bonferroni Correction for six Measures) of subjectively-rated Ability with Performance for all four Tests: The modified Mooney Face Test (*Mooney*), the Glasgow Face Matching Test (*GFMT*), the Cambridge Face Memory Test (*CFMT*), and the Composite Face Test's Holistic Index (*Holistic*) as well as overall Raw Score (*Raw Score*). Also included is the Correlation with *f*. All correlations use the full sample size of N = 397. Confidence intervals at 95% are given between square brackets.

Test	Spearman's ρ	p
Mooney	.21 [.11, .30]	1.5×10^{-4}
GFMT	.29 [.19, .37]	4.4×10^{-8}

TABLE 5

The Spearman Correlation Coefficients and Probability Values (after Bonferroni Correction for six Measures) of subjectively-rated Ability with Performance for all four Tests: The modified Mooney Face Test (*Mooney*), the Glasgow Face Matching Test (*GFMT*), the Cambridge Face Memory Test (*CFMT*), and the Composite Face Test's Holistic Index (*Holistic*) as well as overall Raw Score (*Raw Score*). Also included is the Correlation with *f*. All correlations use the full sample size of N = 397. Confidence intervals at 95% are given between square brackets.

Test	**Spearman's ρ**	***p***
CFMT	.41 [.33, .49]	4.8×10^{-17}
Holistic	.01 [–.09, .11]	1
Composite *(Raw Score)*	.17 [.07, .27]	.004
f	.37 [.28, .45]	1.8×10^{-13}

Autism-Spectrum Quotient. Previous studies have reported a link between Autism-Spectrum Quotient (AQ) and face recognition (Halliday et al., 2014), and since a subset of 316 (203 female) of our 397 participants had previously completed the AQ questionnaire (Baron-Cohen et al., 2001), we also examined this possible link. The mean AQ score in our subset of 316 participants is 17.8 (SD = 7.9), with a range from 3 to 39 (the maximum possible score is 50); a score of 32 or higher is suggestive of autism-spectrum disorder (Baron-Cohen et al., 2001). Though we do not observe a significant sex difference in AQ score (M_{females} = 17.4, M_{males} = 18.7; Mann–Whitney U = 10,002, p = .06), the trend is for males to score higher than females.

When we consider self-rated face-recognition ability, we observe a significant, negative correlation with AQ (Spearman's $\rho = -.23$, $p = 4.8 \times 10^{-5}$, with sex as covariate). However, we do not observe a correlation between AQ and performance for any of our tests, or with the holistic index, or with *f*; and also not when the effect of sex is removed from all variables by means of regression, or when analyses are conducted for females and males separately. Our finding contrasts with that of Halliday and colleagues (2014), who observed a small, but significant, negative correlation between AQ and performance on an immediate memory task using faces ($r = -.20$, $p = .02$, N = 124 university students); we had 89% power to observe an association of the same magnitude ($r^2 = .04$; $\alpha = .008$, corrected for 6 tests). For another population of undergraduate students, Rhodes and colleagues (2013) report correlations between CFMT and AQ that are of opposite sign for men and women. Our own results do not replicate these findings, even when we follow Rhodes and colleagues in calculating a total score (totaling the raw scores of all items, rather than the usual approach of labeling response to items in a binary fashion). It is interesting to note however, that Hedley and colleagues (2011)

found impairment in face recognition only for individuals actually diagnosed with autism, and not as a correlate of autistic traits as measured by the AQ questionnaire.

It could be the case that the relationship between AQ and face cognition does not follow gradually the autistic spectrum, but rather is bimodal, and becomes apparent only when comparing two distinct groups. Indeed, we do observe a significant difference in CFMT performance when comparing participants with AQ of 32 or higher (N = 21, of whom 14 females), to participants with AQ below 32 (N = 295; Mann-Whitney U = 2127.5, p = .02). The latter group scores half a standard deviation higher than the former (M = 75.3% correct vs. 68.9% correct).

Digit ratio. A previous study by Leow and Davies (2012) has linked the face-inversion effect to digit ratio (Manning et al., 1998); and for the present cohort we ourselves have previously reported a significant correlation between digit ratio and performance on our 3AFC adaptation of the Mooney test (Verhallen et al., 2014). However, we do not observe a significant correlation between digit ratio and performance on any of the other three tests, nor with the holistic index, even when the effect of sex is removed. *f* exhibited a small, positive correlation with digit ratio (Spearman's ρ = .12, p = .02) but this correlation would not survive a Bonferroni correction.

Scholastic achievement. We do not hold IQ scores for our participants, but for a subset of our participants (N = 229, of whom 148 were female) we hold self-reported scores for the standard British qualification *General Certificate of Secondary Education* (GCSE; M = 7.45, SD = .67, range = 3.56–8.00[2]), which has been shown to correlate highly with performance on IQ tests (Deary et al. 2007). Neither *f* nor any of the individual face measures showed a significant relationship to GCSE scores (the strongest correlation was with the CFMT: Spearman's ρ = –.12 [–.25, .00], p = .05).

3.4. Genotypic correlates of face-processing ability

We have previously reported a significant genetic association with performance on the Mooney test that we observed in our genome-wide association study (Verhallen et al., 2014). In this study, to allow for multiple comparisons across single-nucleotide

[2] For reference: the distribution of GCSE scores for 2009 (the year prior to the initial PERGENIC test battery) has a mean of 5.08 (SD = 1.73), with a range from 0 to 8 (Stubbs, 2009).

polymorphisms (SNPs), a correction is required for the number of independent genomic locations tested. According to the criterion of Li et al. (2012), a *p*-value of 1.47×10^{-7} is required for an association with any given SNP to achieve significance at $\alpha = .05$ in our study (Verhallen et al., 2014). However, we choose to apply a second rigorous test to guard against false positives: A whole-genome permutation analysis (Purcell et al., 2007).

At the 1.47×10^{-7} level of probability, we observe a further genetic correlate of performance, of ranked overall raw score on the Composite test ($p = 1.31 \times 10^{-7}$; $N = 369$) with rs7701353. This SNP is located in the intergenic region between the genes *BNIP1* and *NKX2–5* on chromosome 5. The minor allele is associated with higher raw score on the Composite test, and the minor allele frequency of rs7701353 is .35 in our sample; the SNP is in Hardy–Weinberg equilibrium ($p = .55$). However, this association does not survive a permutation procedure ($p = .22$; 25,000 permutations), and thus we do not claim it to be significant.

We found no significant genetic associate of *f,* though two SNPs came up as 'suggestive' associations (i.e. associations with an uncorrected probability below 2.95×10^{-6}, but above 1.47×10^{-7}): rs272708 ($p = 1.26 \times 10^{-6}$), which lies on chromosome 7, and rs4866542 ($p = 1.29 \times 10^{-6}$), which lies on chromosome 5 (see also Table 6). These SNPs are both intergenic.

We do not observe any other genetic correlates of performance on the face-processing tests, nor with the holistic index. However, the sample for whom we had genetic information ($N = 370$, of whom 235 female) was small by GWAS standards. For the guidance of other researchers, we record in Table 6 the SNPs that had suggestive associations with our performance measures. Sex was entered as a covariate in all the genetic analyses (for a more detailed description of the genetic methods, see Goodbourn et al., 2014 and Lawrance-Owen et al., 2013).

TABLE 6

Suggestive SNPs for Performance on three out of four Tests: The Mooney Test ('Mooney'), the Cambridge Face Memory Test ('CFMT'), and the Composite Face Test, presented separately for the Holistic Index ('Holistic') and Raw Score ('*Raw Score*'). This Table also lists a suggestive Region for the First Factor of our Factor Analysis ('*f*'). No suggestive SNPs emerged for the GFMT. For each Test we list the suggestive Region, the SNP with the highest Significance Value in that Region ('Lead SNP') along with its Significance Value, the Gene in which the lead SNP is located (or 'intergenic' if it is located in-between Genes), and additional suggestive SNPs in that Region. All suggestive SNPs have *p*-Values below 2.95×10^{-6} and Minor Allele Frequencies above 5%. Performance data for all Measures except the Holistic Index are ranked before being entered into the Genetic Analysis (see also Verhallen et al., 2014). Genomic references were based on the Human February 2009 (GRCh37/hg19) assembly sequence. For further Details of the Genome-Wide Association Analysis, see Verhallen et al., 2014.

Test	Region	Lead SNP	Significance	Gene	Additional SNPs
Mooney	12q24.32	rs9738216	2.09×10^{-7}	*SLC15A4*	rs1059312 rs7962918 rs900982 rs7960920
CFMT	7p15.3	rs272708	1.68×10^{-7}	(intergenic)	
	1q25.1	rs7520814	1.81×10^{-6}	*SLC9C2*	rs16846206
	1p36.21	rs10927998	2.76×10^{-6}	*KAZN*	
	10p12.1	rs7086007	2.89×10^{-6}	*KIAA1217*	rs10508677
Holistic	7q21.13	rs12670363	1.26×10^{-6}	*STEAP2–AS1*	
Raw Score	16q23.1	rs2454141	2.45×10^{-6}	(intergenic)	
f	7p15.3	rs272708	1.26×10^{-6}	(intergenic)	
	5p15.33	rs4866542	1.29×10^{-6}	(intergenic)	

3.5. Absence of a relationship between the holistic index and CFMT performance

The absence of a correlation between the Composite test's holistic index and performance on the CFMT is surprising, since it contradicts previous findings (DeGutis et al., 2013; Richler et al., 2011; Wang et al., 2012). We thus wanted to verify that we had enough power to observe an effect, and to make sure that the relationships between the various conditions (similarity, alignment, congruency) were similar to those reported by previous studies.

The internal reliability of the holistic index from our data is acceptable (.53); together with the internal reliability of the CFMT (.91), the maximum expected correlation is $\sqrt{(.53 \times .91)} = .69$. This is well above the maximum expected correlations

reported in previous studies that did observe a significant correlation between the holistic index and CFMT performance (DeGutis et al., 2013; Ross et al., 2014). Indeed, we *do* observe significant correlations with performance on the CFMT for *d'* of all conditions individually (see Table 7), which is in accordance with previous findings (DeGutis et al., 2013, their Table 1).

TABLE 7

Pearson's *r* and Spearman's ρ, and associated *p*-Values, for Associations between CFMT Performance and *d'* for all Conditions of the Composite Face Test ("Composite") individually: For the *aligned* and *misaligned* Trials, and separately for the *aligned congruent*, *aligned incongruent*, *misaligned congruent*, and *misaligned incongruent* Trials. All correlations use the full sample size of 397. Confidence intervals at 95% are given between square brackets.

	CFMT			
Composite	*Pearson*	*p*	*Spearman*	*p*
Aligned	.36 [.27, .44]	1.15×10^{-13}	.37 [.28, .45]	2.86×10^{-14}
Congruent	.32 [.23, .41]	6.26×10^{-10}	.31 [.22, .40]	2.09×10^{-10}
Incongruent	.26 [.17, .35]	1.53×10^{-7}	.28 [.19, .37]	1.67×10^{-8}
Misaligned	.45 [.37, .53]	3.08×10^{-20}	.44 [.36, .52]	3.42×10^{-20}
Congruent	.43 [.35, .51]	1.88×10^{-18}	.42 [.34, .50]	1.68×10^{-18}
Incongruent	.35 [.26, .43]	1.57×10^{-12}	.36 [.27, .44]	5.44×10^{-14}

Note: *The correlations with CFMT performance—as reported in this table—replicate previous findings (DeGutis et al., 2013), as opposed to the absence of a correlation when we use the holistic index (see main text).*

To investigate further the absence of a correlation between the holistic index and CFMT performance, we look into the internal relationships between our trial variables (similarity, alignment, and congruency) and find them to be consistent with earlier work (e.g. DeGutis et al., 2013; Konar et al., 2010; Richler et al., 2011; Wang et al., 2012). For example, investigating the *same* and *different* trials, we do not observe a significant alignment effect for *different* trials (Wilcoxon signed-rank $W = 36{,}453$, $p_{\text{uncorrected}} = .04$, $r = -.10$), but we do for *same* trials: The mean raw score for *misaligned same* trials is higher than the mean raw score for *aligned same* trials (34.5 vs. 33.1 trials; $W = 21{,}995$, $p = 4.44 \times 10^{-9}$, $r = -.29$). This finding confirms those of Konar et al. (2010), Richler et al. (2011), and Wang et al. (2012).

Separately, we observe a significantly higher mean raw score for *congruent* as compared to *incongruent* trials, regardless of alignment (72.4 vs. 65.3 trials; $W = 70{,}649$, $p = 3.75 \times 10^{-56}$, $r = -.79$). Furthermore, we observe a significant interaction between congruency and alignment (Friedman $\chi^2 = 582.44$, $p = 6.45 \times 10^{-126}$): The mean raw score for *congruent* trials is significantly higher than that for *incongruent* trials, but only when trials are *aligned* ($W = 73{,}010.5$, $p = 7.93 \times 10^{-63}$, $r = -.84$ for *aligned* trials; $W = 32{,}241.5$, $p_{\text{uncorrected}}$

= .03, r = -.11 for *misaligned* trials; see Figure 2). This finding confirms that of DeGutis et al., 2013.

The above findings were virtually identical when using d' instead of raw scores (the effect of *same* and *different* trials cannot be investigated using d', since both *same* and *different* trials are used in calculating d'), and we also obtained very similar results when we performed the analyses using A'—an alternative to d' (Stanislaw and Todorov, 1999).

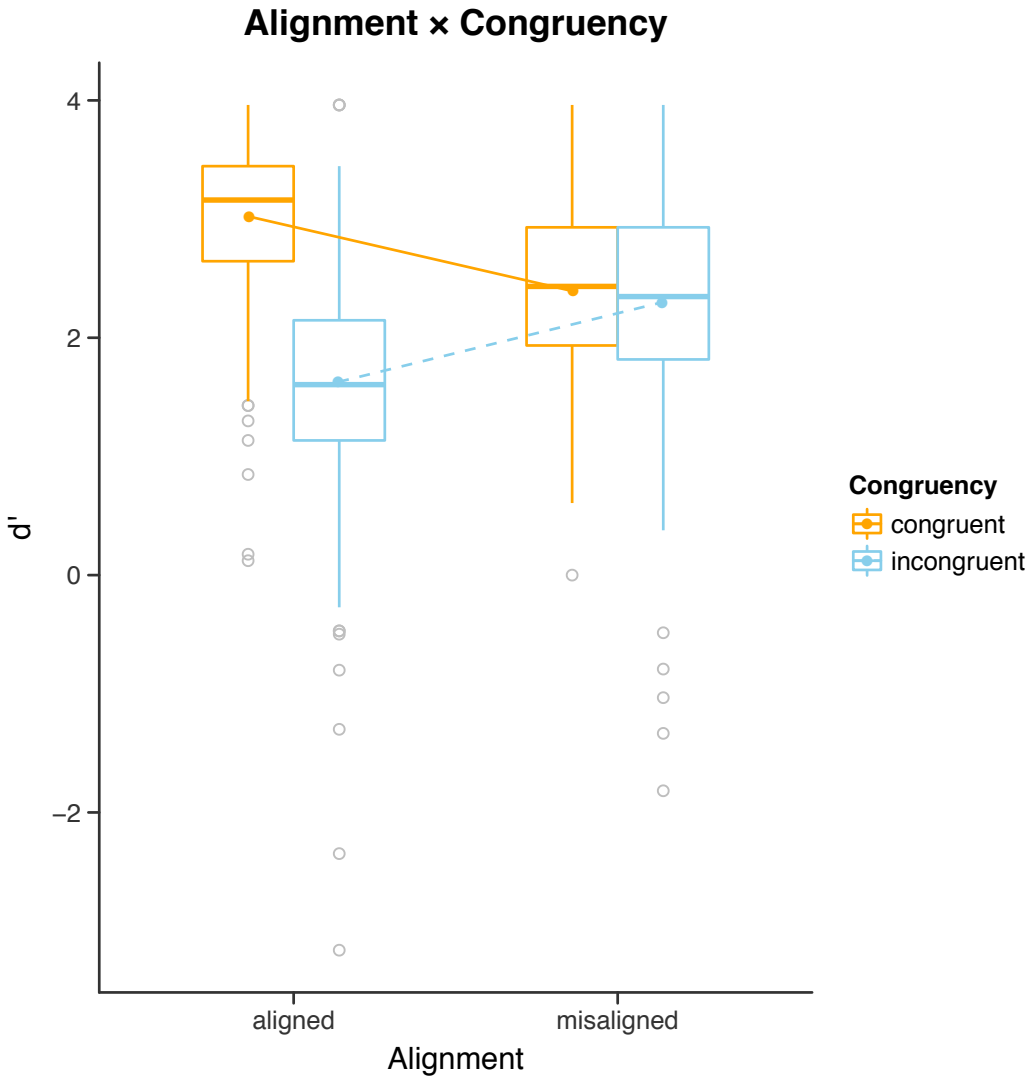

Figure 2. Boxplot illustrating the *alignment* by *congruency* interaction for the Composite Face Test. Mean d' is plotted separately for the four conditions, from left to right: *Aligned congruent*, *aligned incongruent*, *misaligned congruent*, and *misaligned incongruent* trials. Within the boxes, horizontal bars indicate the median, and solid points indicate the mean; the lines connecting the solid points (a solid line for *congruent*, and a dashed line for *incongruent*) illustrate the interaction: Mean d' for *aligned congruent* trials is significantly higher as compared to *aligned incongruent* trials, while mean d' for *misaligned congruent* and *misaligned incongruent* trials do not differ significantly. The bottom and top boundaries of the box indicate the 1st and 3rd quantile, respectively; the whiskers (the vertical lines extending from the bottom and top boundaries of the box) extend to the lowest and highest value that is within 1.5-times the inter-quartile range (IQR) of their respective boundary. Grey open circles are outliers, defined as such by virtue of being 1.5×IQR above or below the 3rd or 1st quantile, respectively. Because data are plotted separately per condition, some of the grey dots denote the same

participant: of the 19 outliers shown here, 17 are individual participants. In all four conditions, *d'* hits ceiling (see also Table 2).

Some previous studies have calculated the holistic index using subtraction rather than regression. The study most similar to ours is that of Richler et al. (2011), whose stimulus set and methods we follow. Those authors calculated a holistic index by subtracting the control condition from the condition of interest. We therefore also compute a holistic index using subtraction, but again we do not observe a significant correlation with performance on the CFMT (Spearman's $\rho = -.05$, $p = .30$; Pearson's $r = -.05$, $p = .36$), whereas Richler and colleagues do (Pearson's $r = .40$, $p = .014$).[3]

For further examples of data exploration, including the exclusion of outliers and use of reaction time instead of accuracy, see the Supplementary Materials (S.3).

[3] In all preceding analyses in this paper we refer to the regression-based holistic index when we write only 'holistic index.'

4. DISCUSSION

4.1. *f*, a general factor underlying the processing of faces

In the field of intelligence testing, a pattern of positive correlations—the 'positive manifold'—is invariably found amongst the diverse tests of a cognitive battery (Mackintosh, 2011). Spearman adopted the term *g* for the common factor that emerges from a factor analysis of test scores. Nevertheless, there are groups of sub-tests that correlate more strongly with each other than they do with other sub-tests; and Thurstone emphasized the specific factors that emerge from a factor analysis.

In so far as intelligence is heritable, the pattern of one general and other specific factors makes good sense. The construction, maintenance and operation of the central nervous system must depend on many thousands of proteins—and the genes that encode them. Most of these genes are polymorphic, either in their coding regions or in the non-coding regions that affect their expression. It is reasonable to suppose that there are many polymorphisms that have a general effect throughout the cerebral cortex, while there will be many others whose effect is limited to particular processing modules.

Just as general and specific factors are observed in the case of intelligence tests, it reasonable to expect general and specific factors underlying the very complex processes that must underlie the discrimination and identification of faces. In the present study, we find highly significant correlations between all pairs of tests, but the correlations differ substantially in their magnitude: the shared variance varies from 4% to 23%. We have proposed the term *f* for the factor on which all the present face tests load, but we emphasize that *f*, like *g*, is no more than a summary of a pattern of correlations and should not be reified. In the case of *g* we now know—from Genome-wide Complex Trait Analysis—that it has a heritability of the order of 50%, but we know equally firmly that it cannot be identified with any single polymorphic gene or even with a small number of genes (Davies et al., 2011; Plomin & Deary, 2015).

We also emphasize that *f* may not be specific to faces. Our results show that several low-level visual functions—contrast sensitivity, recognition of oriented gratings and perception of coherent motion—do not load on this factor; but the possibility remains open that tests of, say, object recognition would load on *f*. Further factor-analytic studies of face and non-face tests offer an attractive route for understanding the nature of *f*.

It is instructive that *f* does not correlate significantly with GCSE scores, our

surrogate measure of *g*. This finding is consistent with earlier studies (using the Cambridge Face Memory Test) that have found little or no correlation between general intelligence and the ability to process faces (Wilmer et al., 2014; Shakeshaft & Plomin, 2015). We must emphasize, however, that a large part of the present sample comprises undergraduate students at a selective university, and is thus restricted in range of intelligence; this would limit our ability to detect any relationship that may be present in a more diverse sample.

Of the four tests of face processing, only the Mooney test loads markedly on the second factor of Table 4. This is the factor on which the two tests of 'coherent form' load very strongly. Perhaps what the three tests have in common is the requirement to integrate local visual features across space. In other words, they perhaps all require the (still-mysterious) process of 'perceptual organization'. However, the detection of coherent motion—which nominally requires similar processes—does not load on this factor, but loads strongly on the third.

4.2. Specific sub-processes in the perception of faces

The four tests of face perception considered here vary in the extent to which they engage different sub-processes required for the perception of faces. Traditional models of the analysis of faces propose two main sub-mechanisms: "structural encoding" and "face recognition units" (Bruce & Young, 1986), or, in another terminology, "early perception of facial features" and "perception of unique identity" (Haxby et al., 2000). Each of these stages, of course, is likely to require many specific sub-stages. Freiwald and Tsao (2010) distinguished six interconnected face-selective regions of the macaque temporal lobe, and identified some of these regions with distinct levels of processing: In the middle lateral and middle fundus patches, neurons were view-specific; in the anterior lateral area, neurons were often tuned to mirror-symmetric views; and in the anterior medial area, neurons were most selective for identity and tended to generalize across many viewpoints.

Let us consider one particularly interesting result from the present study, the low level of shared variance between the Mooney test and the Glasgow Face Matching Test. These results could perhaps be taken to signify that the Mooney test is more a test of closure (a process not required for the GFMT), and that the GFMT is more a test of image comparison (a process not required for the Mooney test). However, the relatively high shared variance (10%) between the Mooney test and the Cambridge Face Memory

Test does suggest that the Mooney test probes sources of variance common to other tests of face processing. For instance, the Mooney test requires the participant to construct—from the two-dimensional, two-tone image—a coherent three-dimensional model both of the light source and of the face (Moore & Cavanagh, 1998); and in performing this feat of internal modeling the participant is likely to draw on stored experiences of faces of different age, sex and demeanor. Many of the required underlying processes will have a less prominent role when a participant performs the Glasgow Face Matching Test, in which the participant is asked to compare only two-dimensional images from similar viewpoints. It may also be relevant that the final phase of the CFMT requires participants to recognize faces in images degraded with random visual noise; it is possible that detection of faces in the two-tone, thresholded Mooney images relies to some extent on the same visual processes as does the extraction of faces embedded in noise.[4]

The Glasgow Face Matching Test and Cambridge Face Memory Test have the highest shared variance: 23%. This is perhaps surprising, because the GFMT primarily entails face *discrimination*, while the CFMT requires face *recognition*; the latter process relies on learning and memory in a way that the former does not. Furthermore, the stimuli used in these two tests differ markedly. Of particular note is that head outline and hair are masked for the CFMT faces, while both are visible in the GFMT; discrimination and recognition performance for unfamiliar faces may rely heavily on such features (Young et al., 1985).

The correlation we observe here between the GFMT and CFMT ($r = .48$) is substantially stronger than the correlation that Burton and colleagues (2010) observe between the GFMT and a custom-made face recognition task ($r = .29$). In fact, Burton and Jenkins (2011) argue that unfamiliar faces are processed as objects rather than faces. If this were indeed the case, then the high shared variance that we observe between a recognition test (CFMT) and an unfamiliar face test (GFMT) could indicate that object-recognition processes are also involved in the recognition of faces (as in the CFMT), or rather that the faces in the CFMT remain effectively unfamiliar. Alternatively, it could be that the 'recognition process' applied during the CFMT involves a 'discrimination process' between the three faces concurrently presented in the CFMT's 3AFC paradigm—a process akin to that used during the GFMT. The high correlation we observe is unlikely to be due to similarity of stimuli between the GFMT and CFMT: The

[4] We thank an anonymous reviewer for this suggestion.

images of the two tests come from different databases, and differ as to whether external features such as face shape and hair are visible. In addition, the low correlation between GFMT and face recognition reported by Burton and colleagues (2010) was observed even though their two tests used images from the same database.

4.3. The holistic index

Despite our large sample size, we do not observe a correlation between the *holistic index* (ostensibly the measure of interest for the Composite test) and performance on any of the other tests, whereas many previous studies report a strong, positive correlation with CFMT performance or a similar measure of face recognition (DeGutis et al., 2013; Richler et al., 2011; Wang et al., 2012). Our results are more in accord with those of Konar et al. (2010), who also do not observe a significant relationship between holistic processing and face identification. However, their task of face identification was arguably more a task of face discrimination (akin to the GFMT), and oddly enough we *do* find a significant correlation between the holistic index and CFMT performance when we pair the idiosyncratic manner in which Konar et al. (2010) calculated the holistic index with a regression-based analysis (see Supplementary Materials, S.3). However, by that point the calculated statistic has become conceptually meaningless. Indeed, most studies that administer the Composite test use either different test versions, or different ways of calculating the holistic index, or both. The comparison of results is thus undermined.

Although Richler and colleagues (2014) have recently developed a new, 3AFC version of the Composite test that could address the aforementioned issues, the holistic index may not reflect a single source of variation: Independently of being good or bad at holistic processing, individuals may vary in the ability to decide actively whether or not to use holistic processing. Indeed, it is interesting to note that d' values for the four conditions separately *do* correlate significantly and strongly with CFMT performance, and that overall raw score on the Composite test correlates significantly (and substantially) with performance on all three other face tests. These correlations suggest that the basic task of judging whether the top halves of two faces are the same or different taps into common face-processing abilities.

It is interesting that we do not find a relationship between Autism-Spectrum Quotient and holistic index, given the extensive evidence that people with autism, and perhaps too some of their relatives, differentially process details at the expense of the perceptual Gestalt (Frith, 2012; Gauthier et al., 2009). However, other studies have

reported intact holistic face processing in autistic individuals (e.g. Cleary et al. 2015; Joseph et al. 2003).

4.4. Self-rating of face-recognition ability

It is striking that a single self-rating of the ability to recognize faces accounts for so much of the variance in CFMT performance (17%). The correlation we observe (r = .41) is slightly higher than a previously reported correlation between CFMT performance and participants' agreement with the statement "I can recognize famous celebrities in photos or on TV" (r = .37; Wilmer et al., 2014). However, our correlation is almost double that obtained when participants judged their ability in comparison to "the average person" (r = .22; Bowles et al., 2009). The latter question might be an external judgment (a question of comparison to an unknown other, thus risking confounds of self-image), whereas our question might tap an internal notion of ability.

4.5. Absence of genetic associations

Although our sample of 370 participants is large by the standards of phenotypic studies, it is small as a genome-wide association study. Thus it may not be remarkable that we fail to identify significant genetic associations of the four face tests in addition to the one association with Mooney performance we have previously reported (Verhallen et al., 2014).

ACKNOWLEDGEMENTS

This work was supported by the Gatsby Charitable Foundation (GAT2903). We thank Mike Burton and colleagues for making available the Glasgow Face Matching Test, Brad Duchaine and colleagues for making available the Cambridge Face Memory Test, and Jennifer Richler and colleagues for making available their version of the Composite Face Test. We are grateful to Ruth Hogg for her part in establishing the PERGENIC battery, and to Emily Clemente, Julien Bauer and Kerry Cliffe of Cambridge Genomic Services for their valuable help.

REFERENCES

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. http://doi.org/10.1023/A:1005653411471

Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., Hogg, R. E., & Mollon, J. D. (2015). A population study of binocular function. *Vision Research*, *110*(PA), 34–50. http://doi.org/10.1016/j.visres.2015.02.017

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., et al. (2009). Diagnosing prosopagnosia: effects of ageing, sex, and participant-stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology*, *26*(5), 423–455. http://doi.org/10.1080/02643290903343149

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327.

Bruce, V., & Young, A. W. (2012). Face perception. London: Psychology Press.

Burton, A. M., & Jenkins, R. (2011). Unfamiliar Face Perception. In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby, *The Oxford Handbook of Face Perception* (pp. 287–306). Oxford: Oxford University Press.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291. http://doi.org/10.3758/BRM.42.1.286

Cleary, L., Brady, N., Fitzgerald, M., & Gallagher, L. (2015). Holistic processing of faces as measured by the Thatcher illusion is intact in autism spectrum disorders. *Autism : the International Journal of Research and Practice*, *19*(4), 451–458. http://doi.org/10.1177/1362361314526005

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., et al. (2011). Genome-wide

association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry*, *16*(10), 996–1005. http://doi.org/10.1038/mp.2011.85

DeGutis, J., Wilmer, J., Mercado, R. J., & Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition*, *126*(1), 87–100. http://doi.org/10.1016/j.cognition.2012.09.004

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21. http://doi.org/10.1016/j.intell.2006.02.001

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. http://doi.org/10.1016/j.neuropsychologia.2005.07.001

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology*, *24*(4), 419–430. http://doi.org/10.1080/02643290701380491

Foreman, N. (1991). Correlates of Performance on the Gollin and Mooney Tests of Visual Closure. *The Journal of General Psychology*, *118*(1), 13–20.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851. http://doi.org/10.1126/science.1194908

Frith, U. (2012). Why we need cognitive explanations of autism. *Quarterly Journal of Experimental Psychology*, *65*(11), 2073–2092. http://doi.org/10.1080/17470218.2012.697178

Gauthier, I., Klaiman, C., & Schultz, R. T. (2009). Face composite effects reveal abnormal face processing in Autism spectrum disorders. *Vision Research*, *49*(4), 470–478. http://doi.org/10.1016/j.visres.2008.12.007

Goodbourn, P. T., Bosten, J. M., Hogg, R. E., Bargary, G., Lawrance-Owen, A. J., & Mollon, J. D. (2012). Do different "magnocellular tasks" probe the same neural substrate? *Proceedings of the Royal Society B: Biological Sciences*, *279*(1745), 4263–4271. http://doi.org/10.1098/rspb.2012.1430

Goodbourn, P. T., Bosten, J. M., Bargary, G., Hogg, R. E., Lawrance-Owen, A. J., & Mollon, J. D. (2014). Variants in the 1q21 risk region are associated with a visual endophenotype of autism and schizophrenia. *Genes, Brain and Behavior*, *13*(2), 144–151. http://doi.org/10.1111/gbb.12096

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282.

Halliday, D. W. R., MacDonald, S. W. S., Sherf, S. K., & Tanaka, J. W. (2014). A reciprocal model of face recognition and autistic traits: evidence from an individual differences perspective. *PloS One*, *9*(5), e94013. http://doi.org/10.1371/journal.pone.0094013

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. http://doi.org/10.1016/S1364-6613(00)01482-0

Hedley, D., Brewer, N., & Young, R. (2011). Face recognition performance of individuals with Asperger syndrome on the Cambridge Face Memory Test. *Autism Research*, *4*(6), 449–455. http://doi.org/10.1002/aur.214

Joseph, R. M., & Tanaka, J. (2003). Holistic and part-based face recognition in children with autism. *Journal of Child Psychology and Psychiatry*, *44*(4), 529–542.

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2010). Holistic Processing Is Not Correlated With Face-Identification Accuracy. *Psychological Science*, *21*(1), 38–43. http://doi.org/10.1177/0956797609356508

Lawrance-Owen, A. J., Bargary, G., Bosten, J. M., Goodbourn, P. T., Hogg, R. E., & Mollon, J. D. (2013). Genetic association suggests that SMOC1 mediates between prenatal sex hormones and digit ratio. *Human Genetics*, *132*(4), 415–421. http://doi.org/10.1007/s00439-012-1259-y

Leow, M. C., & Davis, G. (2012). An index of prenatal steroid exposure predicts adult face perception skills. *Psychonomic Bulletin & Review*, *19*(6), 1094–1100. http://doi.org/10.3758/s13423-012-0317-8

Li, M.-X., Yeung, J. M. Y., Cherny, S. S., & Sham, P. C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human Genetics*, *131*(5), 747–756. http://doi.org/10.1007/s00439-011-1118-2

Mackintosh, N. J. (2011). IQ and Human Intelligence (2nd ed.). Oxford: Oxford University Press.

Manning, J. T., Scutt, D., Wilson, J., & Lewis-Jones, D. I. (1998). The ratio of 2nd to 4th digit length: a predictor of sperm numbers and concentrations of testosterone, luteinizing hormone and oestrogen. *Human Reproduction*, *13*(11), 3000–3004.

Mooney, C. M. (1957a). Age in the development of closure ability in children. *Canadian Journal of Psychology*, *11*(4), 219–226.

Mooney, C. M. (1957b). Closure as affected by configural clarity and contextual consistency. *Canadian Journal of Psychology*, *11*(2), 80–88.

Moore, C., & Cavanagh, P. (1998). Recovery of 3D volume from 2-tone images of novel objects. *Cognition*, *67*(1-2), 45–71.

Peterzell, D. H., & Teller, D. Y. (2000). Spatial frequency tuned covariance channels for red–green and luminance-modulated gratings: psychophysical data from human adults. *Vision Research*, *40*(4), 417–430. http://doi.org/10.1016/s0042-6989(99)00187-x

Plomin, R., & Deary, I. J. (2015). Genetics and intelligence differences: five special findings. *Molecular Psychiatry*, *20*(1), 98–108. http://doi.org/10.1038/mp.2014.105

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. http://doi.org/10.1086/519795

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. http://doi.org/10.1007/s11336-008-9102-z

Rhodes, G., Jeffery, L., Taylor, L., & Ewing, L. (2013). Autistic traits are linked to reduced adaptive coding of face identity and selectively poorer face recognition in men but not women. *Neuropsychologia*, *51*(13), 2702–2708. http://doi.org/10.1016/j.neuropsychologia.2013.08.016

Richler, J. J., & Gauthier, I. (2013). When intuition fails to align with data: A reply to Rossion (2013). *Visual Cognition*, *21*(2), 254–276. http://doi.org/10.1080/13506285.2013.796035

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological Bulletin*, *140*(5), 1281–1302. http://doi.org/10.1037/a0037004

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Holistic processing predicts face recognition. *Psychological Science*, *22*(4), 464–471. http://doi.org/10.1177/0956797611401753

Richler, J. J., Floyd, R. J., & Gauthier, I. (2014). The Vanderbilt Holistic Face Processing Test: A short and reliable measure of holistic face processing. *Journal of Vision*, *14*(11), 10–10. http://doi.org/10.1167/14.11.10

Ross, D. A., Richler, J. J., & Gauthier, I. (2014). Reliability of composite-task measurements of holistic face processing. *Behavior Research Methods*, *47*(3), 736–743. http://doi.org/10.3758/s13428-014-0497-4

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition*, *21*(2), 139–253. http://doi.org/10.1080/13506285.2013.772929

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–257. http://doi.org/10.3758/PBR.16.2.252

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences*, *112*(41), 12887–12892. http://doi.org/10.1073/pnas.1421881112

Spearman, C. E. (1927). The abilities of man, their nature and measurement. London, MacMillan.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.

Stubbs, B. (2009). Student Performance Analysis. Retrieved November 17, 2016, from http://www.bstubbs.co.uk/

Troje, N. F., & Bülthoff, H. H. (1996). Face recognition under varying poses: the role of texture and shape. *Vision Research*, *36*(12), 1761–1771.

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Bargary, G., Lawrance-Owen, A. J., & Mollon, J. D. (2014). An online version of the Mooney Face Test: Phenotypic and genetic associations. *Neuropsychologia*, *63*, 19–25. http://doi.org/10.1016/j.neuropsychologia.2014.08.011

Vigen, M. P., Goebel, R. A., & Embree, L. J. (1982). Adults' performance on a measure of visual closure. *Perceptual and Motor Skills*, *55*(3), 943–952. http://doi.org/10.2466/pms.1982.55.3.943

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual Differences in Holistic Processing Predict Face Recognition Ability. *Psychological Science*, *23*(2), 169–177. http://doi.org/10.1177/0956797611420575

Weigelt, S., Koldewyn, K., & Kanwisher, N. (2012). Face identity recognition in autism spectrum disorders: A review of behavioral studies. *Neuroscience and Biobehavioral Reviews*, *36*(3), 1060–1084. http://doi.org/10.1016/j.neubiorev.2011.12.008

Wilmer, J. B., Germine, L. T., & Nakayama, K. (2014). Face recognition: a model specific ability. *Frontiers in Human Neuroscience*, *8*, 769. http://doi.org/10.3389/fnhum.2014.00769

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive Neuropsychology*, *29*(5-6), 360–392. http://doi.org/10.1080/02643294.2012.753433

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., et al. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, *107*(11), 5238–5241. http://doi.org/10.1073/pnas.0913053107

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A.W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception*, *14(6)*, 737–746.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747–759.

General and specific factors in the processing of faces.

Roeland J Verhallen, Jenny M Bosten, Patrick T Goodbourn,
Adam J Lawrance-Owen,Gary Bargary & J D Mollon

SUPPLEMENTARY MATERIALS

S.1. Calculation of the holistic index

The first step in calculating the *holistic index* is to calculate *d'* from Signal Detection Theory (Stanislaw and Todorov, 1999) for each *alignment* by *congruency* condition (i.e. separately for *aligned-congruent* trials, *aligned-incongruent* trials, *misaligned-congruent* trials, and *misaligned-incongruent* trials). Each of these four conditions has 40 trials: 20 *same* trials and 20 *different* trials (160 trials in total). In the calculation of *d'*—which is done for each condition independently—the hit rate is defined as the total number of "Same" responses to the *same* trials within that condition, divided by the number of these *same* trials; the false-alarm rate is defined as the total number of "Same" responses to *different* trials within that condition, divided by the number of these *different* trials. Thus we have, for each participant, four measures of *d'*, one for each condition. It is important to note that in the calculation of *d'*, hit rates and false-alarm rates are converted to *z*-scores. Should the hit rate or false alarm rate be one or zero, the *z*-score would be infinite, rendering subsequent calculations impossible. Therefore, we applied a *log-linear correction* to both hit rates and false-alarm rates, by adding +.5 to the number of "Same" responses in the numerator, and by adding +1 to the number of trials in the denominator (Hautus, 1995; Stanislaw and Todorov, 1999).

The next step in calculating the *holistic index* is to use the computed *d'* values to calculate the *condition of interest* (by subtracting *d'* of the *aligned incongruent* trials from *d'* of the *aligned congruent* trials) and the *control condition* (by subtracting *d'* of the *misaligned incongruent* trials from *d'* of the *misaligned congruent* trials). Finally, we regress the *control condition* from the *condition of interest*; the residuals from this analysis constitute the *holistic index*.

S.2. Calculation of the split-half reliability for the holistic index

To calculate the split-half reliability of the holistic index, we split the data of the four

conditions by generating—independently for each condition—a random sequence of ten numbers (without replacement) between one and twenty, because there are twenty *same* trials, or hit trials, and twenty *different* trials, or false alarm trials, per condition. The random sequence of ten numbers comprised the ten *same* and ten *different* trials of one subset; the remaining numbers between one and twenty comprised the trials of the other subset. We then calculated—for each subset—d' for the condition of interest and d' for the control condition, and subsequently the holistic index using regression; for both calculations we used the same method as described in the main text. Finally, we calculated the Spearman correlation between the holistic indices of the two subsets: For 5,000 splits of the data, the mean Spearman-Brown corrected correlation was $\rho = .53$ ($SD = .06$).

S.3. Further analysis of the holistic index[1]

The absence of a correlation between the holistic index from the Composite Face Test and performance on the Cambridge Face Memory Test is surprising, since it contradicts previous findings (DeGutis et al., 2013; Richler et al., 2011; Wang et al., 2012). Although d' is a measure free of bias, it still makes certain assumptions about the data: The distributions of hit rates and of false-alarm rates both should be normal, and should exhibit similar standard deviations (Pastore et al., 2003; Stanislaw & Todorov, 1999). We found our distributions (for the Composite test) to be approximately normal, but standard deviation differed between hit rates and false-alarm rates on the aligned-incongruent condition. We thus verified our calculation of d' by calculating A' for all four conditions, and subsequently recalculating the holistic index (Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). We specified hit rate and false-alarm rate in the same way as for the calculation of d', and then calculated—for each condition independently—A' as described by Stanislaw and Todorov (1999, *p.* 142). We used these four measures to calculate the condition of interest and the control condition, and subsequently the holistic index, using the same methods as previously described (see the main text). We did not observe a significant correlation between the holistic index derived from A' and

[1] If only one Spearman statistic is reported, it is always calculated using the holistic index based on the regression method, and reported as the only statistic because it is nearly identical to the result when using the holistic index based on the subtraction method. If Pearson's statistics are not reported, this is because they are nearly identical to the reported Spearman statistics; this is also true for the subtraction method when only one (the regression) statistic is reported.

performance on the CFMT (Spearman's $\rho = -.10$, $p_{\text{uncorrected}} = .05$), nor for that matter with performance on the Mooney Test (Spearman's $\rho = -.10$, $p_{\text{uncorrected}} = .06$) nor with the GFMT (Spearman's $\rho = -.05$, $p_{\text{uncorrected}} = .36$). Our A' and d' holistic indices correlated significantly and positively (Spearman's $\rho = .88$, $p = 2.3 \times 10^{-132}$), and the split-half reliability of the holistic index derived from A' was a Spearman–Brown corrected correlation of $\rho = .47$ ($SD = .08$), the mean of 5,000 iterations. However, since the results with A' are similar to d', and the two correlate significantly, but also because A' might underestimate ability in the presence of bias (Pastore et al., 2003), we will continue to use the holistic index based on d'.

The four studies that investigate the link between the holistic index and face recognition also report the relationships between individual conditions (i.e. *similarity*, *alignment*, and *congruency*) of the Composite test. Additionally, some report the effect of removing outliers, of calculating the holistic index using the partial design, and of using reaction times to calculate the holistic index instead of using accuracy. To further verify and understand our findings, we set out to trace the variety of steps that previous studies had taken, and took a few more of our own.

S.3.1. The Effect of Excluding Outliers

In addition to the analyses described in the main text, and to ensure that our failure to replicate previous findings was not due to outliers, we inspected d' within each of the four conditions, and found a few participants showing values below zero (i.e. a lower hit rate than false alarm rate). We feared these participants might have switched response keys, or that they were judging the change in the bottom halves of the faces rather than the change in the top halves. We thus excluded any participant whose d' values were below zero for any condition, and recalculated the holistic index. Twelve participants were excluded from the analysis; the correlation between the holistic index and CFMT performance remained non-significant (Spearman's $\rho = -.01$, $p = .89$).

An alternative method of spotting participants who potentially were using the wrong strategy, or who had switched response keys, is to omit any participant who responds correctly to fewer than half of the *same incongruent misaligned* trials, or to fewer than half of the *different incongruent misaligned* trials. Since the misaligned condition does not require holistic processing, a consistent response of "Different" on *same incongruent* trials (where the bottom halves are different but the top halves are the same) or of "Same" on

different incongruent trials (where the bottom halves are the same but the top halves are different), would likely be due to a wrong strategy (attending to the bottom halves instead of the top halves), or be due to reversal of the response keys. We excluded any participant who fulfilled either of the two criteria, and recalculated the holistic index. Sixty-one participants were excluded from the analysis; the correlation with CFMT performance remained negligible and non-significant (Spearman's $\rho = -.03$, $p = .51$).

Another method of classifying outliers would be to remove any participant with *d'* values (in any, or multiple, of the four conditions) that were 1.5 times the inter-quartile range above or below the 3rd or 1st quartile, respectively (see the grey, open circles in Figure 2 in the main text). This criterion excluded 17 participants from the analysis; the correlation between holistic index and performance on the CFMT remained negligible and non-significant (Spearman's $\rho = -.03$, $p = .58$).

Since prosopagnosics are thought to be impaired at holistic processing, we investigated whether excluding them from the sample would influence the test statistic. We excluded any participant whose CFMT performance was two standard deviations below the mean (Duchaine and Nakayama, 2006) and recalculated the correlation between CFMT performance and the holistic index: Eleven participants were excluded, and the correlation remained negligible and non-significant (Spearman's $\rho = -.04$, $p = .42$).

S.3.2. Partial Design

Konar et al. (2010), and Wang et al. (2012), both used the partial design, instead of the complete design that is used in the current study and in the studies by Richler et al. (2011) and by DeGutis et al. (2013). The partial design does not have a *congruency* condition, since between the study and target faces the bottom halves are always different. In the terminology of the complete design, the *different* trials of a partial design are always *congruent* (because both the top halves as well as the bottom halves differ between the study face and the target face), whereas the *same* trials are always *incongruent*. Thus, to investigate the partial design using our data from the complete design, we considered only the *same-incongruent* trials and the *different-congruent* trials from both *aligned* and *misaligned* conditions. However, there has been extensive debate about the differential merits of the partial design and the complete design (for details the reader is referred to Rossion, 2013, and the reply to that paper by Richler and Gauthier, 2013). What must be

kept in mind are the possible differences in response bias for the two designs: our 'partial-design measure' that we calculate from our complete-design data could still be influenced by the fact that our trials were part of the complete design (although, see the Supplementary materials in Richler et al., 2011 for a counterargument).

In order to approximate a comparison to previous partial-design studies, we calculated the holistic index for the partial design as follows: d' for the condition of interest was calculated using the response of "Same" on same incongruent-aligned trials as hits, and the response of "Same" on different-congruent aligned trials as false alarms; d' for the control condition was calculated using the response of "Same" on same-incongruent-misaligned trials as hits, and the response of "Same" on different-congruent-misaligned trials as false alarms. We again calculated the holistic index using both the regression method and the subtraction method as detailed above.

Like Konar et al. (2010), we did not observe a significant correlation between performance on the CFMT and the holistic index (based on d') using the partial design and the subtraction method (Spearman's $\rho = -.07$, $p = .14$), and neither when using the regression method (Spearman's $\rho = .09$, $p = .08$). Wang et al. (2012) also did not observe a significant correlation with the holistic index based on accuracy (although they did observe a significant correlation with the holistic index based on reaction time; see the following subsection). However, they calculated their measure of holistic processing in a slightly different way:[2]

$$holistic\ index = \frac{aligned - misaligned}{aligned + misaligned}$$

using only the scores on *same* trials. We also applied this formula to our data, using only the *incongruent* trials in order to replicate the partial design (i.e. we used raw scores on *same-incongruent-aligned* trials for the 'aligned' variable, and raw scores on *same-incongruent-misaligned* trials for the 'misaligned' variable). We did not observe a significant correlation with performance on the CFMT (Spearman's $\rho = -.03$, $p = .53$), confirming the results (Pearson's $r = .03$, $p = .61$) of Wang et al. (2012). Our findings for the holistic index

[2] In their Supplementary material, Wang et al. (2012) describe using an equation in which 'aligned' and 'misaligned' had swapped places (i.e. misaligned – aligned) to calculate the holistic index based on d'. Yet, in their main text, they mention that they used the equation given here for both the holistic index based on reaction times, as well as for the holistic index based on d'. We assume the Supplementary material contained a typographical error, and we will use the equation provided here.

based on *d'* and using the partial design are thus in accordance with the findings of both Konar et al. (2010) and Wang et al. (2012).

Konar et al. (2010) also did not observe a significant correlation between holistic index and face recognition, but although their calculation was almost identical to ours and to those of other previous studies, they report that they swapped the variables within the equation: Instead of subtracting the control condition from the condition of interest (*aligned* – *misaligned*), they subtract the condition of interest from the control condition (*misaligned* – *aligned*). It is unclear whether or not this is a typographical error, since the only effect that swapping the variables would have is a change of the sign of the correlation. However, we felt adventurous and inverted the *regression* calculation of the holistic index using the partial design—regressing the condition of interest from the control condition, i.e. *misaligned* ~ *aligned*, instead of vice versa, as done by DeGutis et al. (2013) and by us in our previous (standard) calculation using *d'* and the regression method. Now, we observed a significant, positive correlation with CFMT performance (Spearman's $\rho = .26$, $p = 1.97 \times 10^{-7}$). However, if we follow current theory, the measure that we obtained here reflects mostly non-holistic processing, as all variation from the condition of interest has been removed. The significant correlation with CFMT performance that we suddenly observe here is thus not with a measure of holistic processing.

S.3.3. Reaction times[3]

In some studies, results were different when using reaction time as compared to accuracy (or *d'* computed from accuracy). For example, Wang et al. (2012) observed a significant, though modest correlation between face-recognition performance and the holistic index based on reaction time (Pearson's $r = .13$, $p < .05$), but did not observe a significant correlation when using the holistic index based on accuracy (Pearson's $r = .03$, $p = .61$), as described above. To compare, we calculated the holistic index based on reaction time,[4] using the formula from Wang and colleagues, which uses the partial design: (*aligned* – *misaligned*) divided by (*aligned* + *misaligned*). As with accuracy, we did not observe a

[3] All trials are taken into account for calculations using reaction time, not just the trials on which the participant scored correctly.

[4] As opposed to the calculation of *d'*, where *same* and *different* trials are used for hits and false alarms, the calculation of reaction time simply pools the two types of trials. This results in four reaction-time averages for each of the four conditions. However, owing to a technical error we have reaction times only for the first 80 trials (out of 160 trials in total).

significant correlation with CFMT performance (Spearman's ρ = –.03, p = .58; Pearson's r = .08, p = .09).

Konar et al. (2010) did not find a significant correlation for accuracy, or for reaction time. When we used their formula of aligned minus misaligned trials, we did not observe a significant correlation with reaction time (Spearman's ρ = .12, $p_{\text{uncorrected}}$ = .02; Pearson's r = .08, p = .09). However, as with accuracy, we did find a significant though marginal correlation when using regression, instead of subtraction, in this formula (Spearman's ρ = .15, p = .003; Pearson's r = .13, p = .01). Unfortunately, Konar et al. (2010) do not report results using the regression method, thus we cannot compare our results to theirs.

Finally, the formula we had used to calculate the holistic index based on d' (see the main text), we also applied to reaction time: Mean reaction time for *aligned-congruent* trials minus mean reaction time for *aligned-incongruent* trials constituted the condition of interest, while mean reaction time for *misaligned-congruent* trials minus mean reaction time for *misaligned-incongruent* trials constituted the control condition. Subsequently, subtracting the control condition from the condition of interest, or regressing the control condition from the condition of interest and taking the residuals, constituted the holistic index for the subtraction and regression methods, respectively. Using this formula, we did not observe a significant correlation between performance on the CFMT and the holistic index based on reaction time when using the regression method (Spearman's ρ = –.10, $p_{\text{uncorrected}}$ = .04; Pearson's r = –.06, p = .26), nor when using the subtraction method (Spearman's ρ = –.03 , p = .55; Pearson's r = –.07, p = .14).

For this last holistic index based on reaction time, we calculated the Spearman–Brown corrected split-half reliability in the same way as we had previously done for the holistic index based on d' (see §S.1). With 5,000 splits of the data, mean ρ was .26 (SD = .08). The upper-bound correlation with CFMT performance was thus $\sqrt{.26 \times .91} = .49$. The reliability of this measure was low, because we had reaction times only for the first half of our test (trials 1 to 80, out of 160 trials in total) owing to a technical error; though reliability remains slightly higher than the reliability of .43 reported by Wang et al. (2012), who did observe a significant effect.

S.3.4. Conclusion

Although our data contained a non-trivial ceiling effect (especially for the *aligned-congruent* trials) we consistently did not observe a correlation between the holistic index based on d', and face recognition as indexed by the CFMT, using testing methods as well as statistical analyses that in previous studies resulted in significant correlations. There were only two instances in which we did observe a significant correlation with CFMT performance, both of which used a combination of calculations not previously reported: one, when using the idiosyncratic formula of Konar et al. (2010) for d' from the partial design, but substituting subtraction with regression (Spearman's $\rho = .26$, $p = 2.0 \times 10^{-7}$); and two, when using the idiosyncratic formula of Konar et al. (2010) for *reaction times* from the partial design, but again substituting subtraction with regression (Spearman's $\rho = .15$, $p = .003$).

Our investigation into the various methods of data analysis highlights the complexity of the current holistic-processing literature: The contradictory results of the various studies seem to be confounded by numerous differences between them. Not only are different stimulus sets and different experimental methodologies used for both the holistic index (the complete vs. the partial design) as well as for the measure of face recognition (the CFMT vs. other, unstandardized tests), but the methods of data analysis differ as well: The holistic index has been calculated using subtraction, regression, or even a custom calculation such as Wang et al., 2012; and the exact sequence of conditions within these calculations (whether the control condition is subtracted or regressed from the condition of interest, or vice versa) has been varied. With stimuli, experimental methods, and methods of data analysis that are all unstandardized, it remains very difficult to draw firm and equivocal conclusions as to how strong (or weak) the relationship between holistic processing and face recognition ability might be.

S.4. References

DeGutis, J., Wilmer, J., Mercado, R. J., & Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition*, *126*(1), 87–100. http://doi.org/10.1016/j.cognition.2012.09.004

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. http://doi.org/10.1016/j.neuropsychologia.2005.07.001

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of

d′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. http://doi.org/10.3758/BF03203619

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2010). Holistic Processing Is Not Correlated With Face-Identification Accuracy. *Psychological Science*, *21*(1), 38–43. http://doi.org/10.1177/0956797609356508

Macmillan, N. A., & Creelman, C. D. (2005). Detection theory: a user's guide (2nd ed.). London: Lawrence Erlbaum Associates.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, *10*(3), 556–569. http://doi.org/10.3758/BF03196517

Richler, J. J., & Gauthier, I. (2013). When intuition fails to align with data: A reply to Rossion (2013). *Visual Cognition*, *21*(2), 254–276. http://doi.org/10.1080/13506285.2013.796035

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Holistic processing predicts face recognition. *Psychological Science*, *22*(4), 464–471. http://doi.org/10.1177/0956797611401753

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition*, *21*(2), 139–253. http://doi.org/10.1080/13506285.2013.772929

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual Differences in Holistic Processing Predict Face Recognition Ability. *Psychological Science*, *23*(2), 169–177. http://doi.org/10.1177/0956797611420575